In: Richesson RL, Andrews JE, editors. Clinical research informatics. London: Springer-Verlag; 2018. p. 313-339.

Knowledge Representation and Ontologies

Kin Wah Fung and Olivier Bodenreider

National Library of Medicine, Bethesda, Maryland, USA

Abstract

The representation of medical data and knowledge is fundamental in the field of medical informatics. Ontologies and related artifacts are important tools in knowledge representation, yet they are often given little attention and taken for granted. In this chapter, we give an overview of the development of medical ontologies, including available ontology repositories and tools. We highlight some ontologies that are particularly relevant to clinical research, and describe with examples the benefits of using ontologies to facilitate research workflow management, data integration and electronic phenotyping.

Keywords

Knowledge representation, Biomedical ontologies, Research metadata ontology, Data content ontology, Ontology-driven knowledge bases, Data integration, Electronic phenotyping

Ontologies have become important tools in biomedicine, supporting critical aspects of both health care and biomedical research, including clinical research [1]. Some even see ontologies as integral to science [2]. Unlike terminologies (focusing on naming) and classification systems (developed for partitioning a domain), ontologies define the types of entities that exist, as well as their interrelations. And while knowledge bases generally integrate both definitional and assertional knowledge, ontologies focus on what is always true of entities, i.e., definitional knowledge [3]. In practice, however, there is no sharp distinction between these kinds of artifacts and 'ontology' has become a generic name for a variety of knowledge sources with important differences in their degree of formality, coverage, richness and computability [4].

Ontology development

Ontology development has not yet been formalized to the same extent as, say, database development has, and there is still no equivalent for ontologies to the entity-relationship model. However, ontology development is guided by fundamental ontological distinctions and supported by the formalisms and tools for knowledge representation that have emerged over the past decades. Several top-level ontologies provide useful constraints for the development of domain ontologies and one of the most recent trends is increased collaboration among the creators of ontologies for coordinated development.

Important ontological distinctions

A small number of ontological distinctions inherited from philosophical ontology provide a useful framework for creating ontologies. The first distinction is between types and instances. Instances correspond to individual entities (e.g., my left kidney, the patient identified by 1234), while types represent the common characteristics of sets of instances (e.g., a *kidney* is a bean-shaped, intra-abdominal organ – properties common to all kidneys) [5]. Instances are related to the corresponding types by the relation ***instance of***. For example, my left kidney is an ***instance of*** *kidney*. (It must be noted that most biomedical ontologies only represent types in reference to which the corresponding instances are recorded in patient records and in laboratory notebooks). Another fundamental distinction is between continuants and occurrents [6]. While continuants exist (endure) through time, occurrents go through time in phases. Roughly speaking, objects (e.g., a liver, an endoscope) are continuants and processes (e.g., the flow of blood through the mitral valve) are continuants. One final distinction is made between independent and dependent continuants. While the kidney and its shape are both continuants, the shape of the kidney "owes" its existence to the kidney (i.e., there cannot be a kidney shape unless there is a kidney in the first place). Therefore, the kidney is an independent continuant (as most objects are), whereas its shape is a dependent continuant (as are qualities, functions and dispositions, all dependent on their bearers). These distinctions are important for ontology developers, because they help organize entities in the ontology and contribute to consistent ontology development, both within and, more importantly for interoperability, across ontologies.

Building blocks: Top-level ontologies and Relation Ontology

These ontological distinctions are so fundamental that they are embodied by top-level ontologies such as BFO [7] (Basic Formal Ontology) and DOLCE [8] (Descriptive Ontology for Linguistic and Cognitive Engineering). Such upper-level ontologies are often used as building blocks for the development of domain ontologies. Instead of organizing the main categories of entities of a given domain under some artificial root, these categories can be implemented as specializations of types from the upper-level ontology. For example, a protein is an independent continuant, the catalytic function of enzymes is a dependent continuant, and the activation of an enzyme through phosphorylation is an occurrent. Of note, even when they do not leverage an upper-level ontology, most ontologies implement these fundamental distinctions in some way. For example, the first distinction made among the semantic types in the UMLS Semantic Network [9] is between *Entity* and *Event*, roughly equivalent to the distinction between continuants and occurrents in BFO. While BFO and DOLCE are generic upper-level ontologies, Bio-Top – itself informed by BFO and DOLCE – is specific to the biomedical domain and provides types directly relevant to this domain, such as *Chain Of Nucleotide Monomers* and *Organ System*. BFO forms the backbone of several ontologies form the Open Biomedical Ontologies (OBO) family and Bio-Top has also been reused by several ontologies. Some also consider the UMLS Semantic Network, created for categorizing concepts from the UMLS Metathesaurus, an upper-level ontology for the biomedical domain [9].

In addition to the ontological template provided for types by upper-level ontologies, standard relations constitute an important building block for ontology development and help ensure consistency across ontologies. The small set of relations defined collaboratively in the Relation Ontology [5], including ***instance of***, ***part of*** and ***located in***, has been widely reused.

Formalisms and tools for knowledge representation

Many ontologies use description logics for their representation. Description logics (DLs) are a family of knowledge representation languages, with different levels of expressiveness [10]. The main advantage of using DL for ontology development

is that DL allows developers to test the logical consistency of their ontology. This is particularly important for large biomedical ontologies. Ontologies including OCRe, OBI, SNOMED CT and the NCI Thesaurus, discussed later in this chapter, all rely on some sort of DL for their development.

Ontologies are key enabling resources for the Semantic Web, the "web of data", where resources annotated in reference to ontologies can be processed and linked automatically [11]. It is therefore not surprising that the main language for representing ontologies, OWL – the Web Ontology Language, has its origins in the Semantic Web. OWL is developed under the auspices of the World Wide Web Consortium (W3C). The current version of the OWL specification is OWL 2, which offers several profiles (sublanguages) corresponding to different levels of expressivity and support of DL languages [12]. Other Semantic Web technologies, such as RDF/S (Resource Description Framework Schema) [13] and SKOS (Simple Knowledge Organization System) [14] have also been used for representing taxonomies and thesauri, respectively.

The OWL syntax can be overwhelming to biologists and clinicians, who simply want to create an explicit specification of the knowledge in their domain. The developers of the Gene Ontology created a simple syntax later adopted for the development of many ontologies from the Open Biomedical Ontologies (OBO) family. The so-called OBO syntax [15, 16] provides an alternative to OWL, to which it can be converted [17].

The most popular ontology editor is Protégé, developed at the Stanford Center for Biomedical Informatics Research for two decades [18, 19]. Originally created for editing frame-based ontologies, Protégé now supports OWL and other Semantic Web languages. Dozens of user-contributed plugins extend the standalone version (e.g., for visualization, reasoning services, support for specific data formats) and the recently-developed web version of Protégé supports the collaborative development of ontologies. Originally created to support the development of the Gene Ontology, OBO-Edit now serves as a general ontology editor [20, 21]. Simpler than Protégé, OBO-Edit has been used to develop many of the ontologies from the Open Biomedical Ontologies (OBO) family. Rather than OWL, OBO-

Edit uses a specific format, the OBO syntax, for representing ontologies. Both Protégé and OBO-Edit are open-source, platform independent software tools.

OBO Foundry and other harmonization efforts

Two major issues with biomedical ontologies are their proliferation and their lack of interoperability. There are several hundreds of ontologies available in the domain of life sciences, some of which overlap partially but do not systematically cross-reference equivalent entities in other ontologies. The existence of multiple representations for the same entity makes it difficult for ontology users to select the right ontology for a given purpose and requires the development of mappings between ontologies to ensure interoperability. Two recent initiatives have offered different solutions to address the issue of uncoordinated development of ontologies.

The OBO Foundry is an initiative of the Open Biomedical Ontologies (OBO) consortium, which provides guidelines and serves as coordinating authority for the prospective development of ontologies [22]. Starting with the Gene Ontology, the OBO Foundry has identified kinds of entities for which ontologies are needed and have selected candidate ontologies to cover a given subdomain, based on a number of criteria. Granularity and fundamental ontological distinctions form the basis for identifying subdomains. For example, independent continuants (entities) at the molecular level include proteins (covered by the protein ontology), while macroscopic anatomical structures are covered by the Foundational Model of Anatomy. In addition to syntax, versioning and documentation requirements, the OBO Foundry guidelines prescribe that OBO Foundry ontologies be limited in scope to a given subdomain and orthogonal. This means, for example, that an ontology of diseases referring to anatomical structures as the location of diseases (e.g., *mitral valve regurgitation* ***has location*** *mitral valve*) should cross-reference entities from the reference ontology for this domain (e.g., the Foundational Model of Anatomy for *mitral valve*), rather than redefine these entities. While well adapted to coordinating the prospective development of ontologies, this approach is extremely prescriptive and virtually excludes the many legacy ontologies used in the clinical domain, including SNOMED CT and the NCI Thesaurus.

The need for harmonization, i.e., making existing ontologies interoperable and avoiding duplication of development effort, has not escaped the developers of large clinical ontologies. SNOMED International, in charge of the development of SNOMED CT, is leading a similar harmonization effort in order to increase interoperability and coordinate the evolution of legacy ontologies and terminologies, including Logical Observation Identifiers Names and Codes (LOINC, for laboratory and clinical observations), the International Classification of Diseases (ICD), Orphanet (for rare diseases), the Global Medical Device Nomenclature Agency (for medical devices), and the International Classification for Nursing Practice (ICNP, for nursing diagnoses) [23].

Ontologies of particular relevance to clinical research

Broadly speaking, clinical research ontologies can be classified into those that model the characteristics (or metadata) of the clinical research and those that model the data contents generated as a result of the research. [24] Research metadata ontologies center around characteristics like study design, operational protocol and methods of data analysis. They define the terminology and semantics necessary for formal representation of the research activity and aim to facilitate activities such as automated management of clinical trials and cross-study queries based on study design, intervention or outcome characteristics. Ontologies of data content focus on explicitly representing the information model of and data elements (e.g. clinical observations, laboratory test results) collected by the research, with the aim to achieve data standardization and semantic data interoperability. Important examples of the two types of ontology will be described in more detail.

Research metadata ontology

We did a survey of the public repositories of ontologies in the Open Biomedical Ontologies (OBO) library hosted by the National Center of Biomedical Ontology and the FAIRsharing.org website hosted by the University of Oxford. [25, 26] We used the keywords “clinical trial”, “research” and “investigation” for searching. We learned about the identified research metadata ontologies through their online information and literature search. Ontologies with little or no available

information and evidence of ongoing use are not included here. We found three ontologies that are actively maintained and used: Ontology of Clinical Research (OCRe), Ontology for Biomedical Investigations (OBI) and Biomedical Research Integrated Domain Group model ontology (BRIDG).

Ontology of Clinical Research

The primary aim of OCRe is to support the annotation and indexing of human studies to enable cross-study comparison and synthesis. [27, 28] Developed as part of the Trial Bank Project, OCRe provides terms and relationships for characterizing the essential design and analysis elements of clinical studies. Domain specific concepts are covered by reference to external vocabularies. Workflow related characteristics (e.g. schedule of activities) and data structure specification (e.g. schema of data elements) are not within the scope of OCRe.

The three core modules of OCRe are:

1. Clinical module – the upper-level entities (e.g. clinician, study subject)
2. Study design module –models study design characteristics (e.g. investigator assigned intervention, external control group)
3. Research module – terms and relationships to characterize a study (e.g. outcome phenomenon, assessment method)

OCRe entities are mapped to the Basic Formal Ontology (BFO).

Ontology for Biomedical Investigations

Unlike OCRe which is rooted in clinical research, the origin of OBI is in the molecular biology research domain. [29, 30] The forerunner of OBI is the MGED Ontology developed by the Microarray Gene Expression Data Society for annotating microarray data. Through collaboration with other groups in the 'OMICS' arena such as the Proteomics Standards Initiative (PSI) and Metabolomics Standards Initiative (MSI), MGED Ontology was expanded to cover proteomics and metabolomics and was subsequently renamed Functional Genomics Investigation Ontology (FuGO). [31] The scope of FuGO was later extended to cover clinical and epidemiological research and biomedical imaging,

resulting in the creation of OBI, which aims to cover all biomedical investigations [32].

As OBI is an international, cross-domain initiative, the OBI Consortium draws upon a pool of experts from many fields, including even fields outside biology such as environmental science and robotics. The goal of OBI is to build an integrated ontology to support the description and annotation of biological and clinical investigations, regardless of the particular field of study. OBI also uses the BFO as its upper-level ontology and all OBI classes are a subclass of some BFO class. OBI covers all phases of the experimental process, and the entities or concepts involved, such as study designs, protocols, instrumentation, biological material, collected data and their analyses. OBI also represents roles and functions which can be used to characterize and relate these entities or concepts.
Specifically, OBI covers the following areas:

1. Biological material – e.g. blood plasma
2. Instrument – e.g. microarray, centrifuge
3. Information content – e.g. electronic medical record, biomedical image
4. Design and execution of an investigation – e.g. study design, electrophoresis
5. Data transformation – e.g. principal components analysis, mean calculation

For domain-specific entities, OBI makes reference to other ontologies such as Gene Ontology (GO) and Chemical Entities of Biological Interest (ChEBI). The ability of OBI to adequately represent and integrate different biological experimental processes and their components has been demonstrated in examples from several domains, including neuroscience and vaccination.

Biomedical Research Integrated Domain Group (BRIDG) Model Ontology

The Biomedical Research Integrated Domain Group (BRIDG) Model is a collaborative effort engaging stakeholders from the Clinical Data Interchange Standards Consortium (CDISC, described in more detail in a later chapter), HL7 Regulated Clinical Research Information Management Technical Committee (RCRIM TC), National Cancer Institute (NCI) and US Food and Drug Administration (FDA). [33-35]The goal of the BRIDG Model is to produce a

shared view of the dynamic and static semantics for the domain of protocol-driven research and its associated regulatory artifacts, defined as "the data, organization, resources, rules, and processes involved in the formal assessment of the utility, impact, or other pharmacological, physiological, or psychological effects of a drug, procedure, process, or device on a human, animal, or other subject or substance plus all associated regulatory artifacts required for or derived from this effort, including data specifically associated with post-marketing adverse event reporting".

One important function of the BRIDG model is to facilitate integration and meaningful data exchange from biological, translational and clinical studies with data from health systems by providing a common understanding of biomedical research concepts and their relationships with health care semantics.

The BRIDG model (version 5.0) is divided into 9 subdomains:

1. Common – concepts and semantics shared across different types of protocol-driven research e.g., people, organizations, places, materials
2. Protocol representation – planning and design of a clinical research protocol e.g., study objective, outcome measure, inclusion criteria
3. Study conduct – concepts related to execution of a research protocol e.g., study site investigator, funding source, specimen collection
4. Adverse events – safety-related activities such as detection, evaluation and follow-up reporting of adverse events
5. Statistical analysis – planning and performance of the statistical analysis of data collected during execution of the protocol
6. Experiment – design, planning, resourcing and execution of biomedical experiments e.g., devices and parameters, variables that can be manipulated
7. Biospecimen – collection and management of biospecimens
8. Molecular biology – including genomics, transcriptomics, proteomics, pathways, biomarkers and other concepts
9. Imaging – covers imaging semantics such as image acquisition, processing and reconstruction

The experiment, biospecimen and molecular biology subdomains are introduced since version 4.0 in response to calls for BRIDG to support molecular-based medicine, in which treatment of disease is informed by the patient's genome and other molecular characteristics. The imaging subdomain is new for version 5.0 to facilitate interfacing between a clinical trials management system and imaging systems. To enhance interoperability with other ongoing data modeling efforts, the BRIDG model has been mapped to the Common Data Model (CDM) of the Observational Health Data Sciences and Informatics (OHDSI) network. Ability to map to other clinical trial ontologies has also been demonstrated. [36] One of the future priorities of BRIDG is vocabulary binding. Historically, BRIDG is ontology and terminology agnostic and no formal binding is provided between vocabularies and classes and attributes within BRIDG. Recognizing the value of improving semantic interoperability, future work will bind BRIDG class attributes to one or more common terminologies from medicine and research.

The BRIDG model is available as an OWL ontology. It is also available as a UML representation (intended for domain experts and architects), and as an HL7 reference information model (RIM) representation in Visio files.

Data content ontology

While there are relatively few research metadata ontologies, there is a myriad of ontologies that cover research data contents. Unlike metadata ontologies, in this group the distinction between ontologies, vocabularies, classifications and code sets often gets blurred, and we shall refer to all of them as "terminologies". As clinical research is increasingly conducted based on EHR data (e.g., pragmatic trials), the separation between terminologies for clinical research and healthcare is also becoming less important. We have chosen several terminologies for more detailed discussion here because of their role in clinical research and in electronic health records. These terminologies are: the National Cancer Institute Thesaurus (NCIT), Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT), Logical Observation Identifiers Names and Codes (LOINC), RxNorm, International Classifications of Diseases (ICD) and Current Procedural

Terminology (CPT). All these terminologies are available through the Unified Medical Language System (UMLS) and the BioPortal ontology repositories (see below).

National Cancer Institute Thesaurus (NCIT)

NCIT is developed by the U.S. National Cancer Institute (NCI). It arose initially from the need for an institution-wide common terminology to facilitate interoperability and data sharing by the various components of NCI. [37-39] NCIT covers clinical and basic sciences as well as administrative areas. Even though the content is primarily cancer-centric, since cancer research spans a broad area of biology and medicine, NCIT can potentially serve the needs of other research communities. Due to its coverage of both basic and clinical research, NCIT is well positioned to support translational research. NCIT was the reference terminology for the NCI's Cancer Biomedical Informatics Grid (caBIG) and other related projects. It was one of the U.S. Federal standard terminologies designated by the Consolidated Health Informatics (CHI) initiative and it hosts many CDISC concepts and value sets.

NCIT contains about 120,000 concepts organized into 19 disjoint domains. A concept is allowed to have multiple parents within a domain. NCIT covers the following areas:

1. Neoplastic and other diseases
2. Findings and abnormalities
3. Anatomy, tissues and subcellular structures
4. Agents, drugs and chemicals
5. Genes, gene products and biological processes
6. Animal models of disease
7. Research techniques, equipment and administration

NCIT is updated monthly. It is in the public domain under an open content license and is distributed by the NCI in OWL format.

SNOMED Clinical Terms (SNOMED CT)

SNOMED CT was originally developed by the College of American Pathologists. Its ownership was transferred to SNOMED International (originally called

International Health Terminology Standards Development Organisation, IHTSDO) in 2007 to enhance international governance and adoption. [40] SNOMED CT has been steadily gaining momentum as the emerging international standard clinical terminology. The number of member countries of SNOMED International has more than tripled since its inception. There are currently 33 member countries including U.S, United Kingdom, Canada, Australia, India, Malaysia, Netherlands, Sweden and Spain. SNOMED CT is used in over 50 countries in the world. [41] SNOMED CT is the most comprehensive clinical terminology available today, with over 340,000 active concepts. The concepts are organized into 19 disjoint hierarchies. Within each hierarchy, a concept is allowed to have multiple parents. Additionally, SNOMED CT provides a rich set of associated relations (across hierarchies), which form the basis for the logical definitions of its concepts. The principal use of SNOMED CT is to encode clinical information (e.g. diseases, findings, procedures). It also has comprehensive coverage of drugs, organisms and anatomy. SNOMED CT is a designated terminology for the problem list, procedures and other data fields according to the Meaningful Use of EHR incentive program of the U.S. Centers for Medicare & Medicaid Services (CMS) [42, 43]. After the Meaningful Use program ended in 2017, the requirements for SNOMED CT use persist in the subsequent Merit-based Incentive Payment System (MIPS) and Promoting Interoperability programs.

SNOMED CT is updated twice every year. The use of SNOMED CT is free in all SNOMED International member countries, in low-income countries as defined by the World Bank, and for qualified research projects in any country. SNOMED CT is distributed by the National Release Center of the SNOMED International member countries.

Logical Observation Identifiers, Names and Codes (LOINC)

LOINC is developed by the Regenstrief Institute, a nonprofit biomedical informatics and healthcare research organization associated with Indiana University. [44] LOINC's primary role is to provide identifiers and names for laboratory and clinical observations that will facilitate the unambiguous exchange and aggregation of clinical results for many purposes, including care delivery,

quality assessment, public health and research purposes. [45] The laboratory section of LOINC covers the usual categories in clinical laboratory testing such as chemistry, urinalysis, hematology, microbiology, molecular genetics and others. This section accounts for about two-thirds of LOINC codes. The clinical section covers a very broad scope, from clinical documents, anthropomorphic measures to cardiac and obstetrical ultrasound. Each LOINC code corresponds to a single kind of observation, measurement or test result. A LOINC term includes six parts: component, kind of property, time aspect, system, type of scale and type of method (optional). LOINC has over 80,000 terms. In 2013, the Regenstrief Institute and SNOMED International formed a long-term collaborative relationship with the objective of developing coded content to support order entry and result reporting by linking SNOMED CT and LOINC. This landmark agreement aims to reduce duplication of effort and provide a common framework within which to use the two terminologies.

In the U.S., LOINC has been adopted by large reference laboratories, health information exchanges, health care organizations and insurance companies. LOINC is also a designated terminology for the EHR under the Meaningful Use program. Internationally, LOINC has over 60,000 registered users from 172 countries. At least 15 countries have chosen LOINC as a national standard. LOINC is updated twice a year. Use of LOINC is free upon agreeing to the terms-of-use in the license.

RxNorm

RxNorm is a standard nomenclature for medications developed by NLM. [46] RxNorm provides normalized names for clinical drugs and links its names to many of the drug vocabularies commonly used in pharmacy management and drug interaction software, including those of First Databank, Micromedex, Gold Standard Drug Database, and Multum. RxNorm also integrates drugs from sources like DrugBank and the Anatomical Therapeutic Chemical (ATC) drug classification system, often used in research projects. By providing links between these vocabularies, RxNorm can mediate messages between systems not using the same software and vocabulary. The focus of RxNorm is at the clinical drug level, represented as a combination of ingredients, strength and dose form. The clinical

drug is linked by semantic relationships to other drug entities such as ingredients and drug packs. Non-therapeutic radiopharmaceuticals, bulk powders, contrast media, food, dietary supplements, and medical devices (e.g., bandages and crutches) are all out of scope for RxNorm. RxNorm has about 37,000 generic clinical drugs, 22,000 branded clinical drugs and 11,000 ingredients. The Current Prescribable Content Subset is a subset of currently prescribable drugs in RxNorm. The subset is intended to be an approximation of the prescription drugs currently marketed in the U.S., and it also includes some frequently-prescribed over-the-counter drugs.

RxNorm is the designated terminology for medications and medication allergies according to the Meaningful Use incentive program. The Centers for Medicare and Medicaid Services (CMS) uses RxNorm in its Formulary Reference File and to define the value sets for clinical quality measures. The National Council for Prescription Drug Programs (NCPDP) uses RxNorm in its SCRIPT e-prescribing and Formulary and Benefit standards. The Department of Veterans Affairs (VA) and the Department of Defense (DoD) use RxNorm to enable bi-directional real-time data exchange for medication and drug allergy information. [47]

RxNorm is released as a full data set every month. There are weekly updates for newly-approved drugs. To download the RxNorm files, a UMLS user license is required because some RxNorm content comes from commercial drug knowledge sources and is proprietary.

International Classification of Disease (ICD)

The root of ICD can be traced back to the International List of Causes of Death created 150 years ago. [48] ICD is endorsed by the World Health Organization (WHO) to be the international standard diagnostic classification for epidemiology, health management and clinical purposes. The current version of ICD is ICD-10 which was first published in 1992. ICD-11 is still under development. Apart from reporting national mortality and morbidity statistics to WHO, many countries use ICD-10 for reimbursement and healthcare resource allocation. To better suit their national needs, several countries have created national extensions to ICD-10, including ICD-10-AM (Australia), ICD-10-CA (Canada) and ICD-10-CM (U.S.).

In the U.S., ICD-9-CM was used until 2015 and was replaced by ICD-10-CM. Because of the requirement of ICD codes for reimbursement, they are ubiquitous in the EHR and insurance claims data. There is a four-fold increase in the number of codes from ICD-9-CM to ICD-10-CM, due to the more granular disease codes and capture of additional healthcare dimensions (e.g., episode of encounter, stage of pregnancy). [49] CMS provides forward and backward maps between ICD-9-CM and ICD-10-CM, which are called General Equivalence Maps (GEMs). These maps are useful for conversion of coded data between the two versions of ICD. [50]

While ICD-9-CM covers both diagnosis and procedures, ICD-10-CM does not cover procedures. A brand-new procedure coding system called ICD-10-PCS was developed by CMS to replace the ICD-9-CM procedure codes for reporting of inpatient procedures. [51] ICD-10-PCS is a radical departure from ICD-9-CM and uses a multi-axial structure. Each ICD-10-PCS code has seven digits, each covering one aspect of a procedure such as body part, root operation, approach and device. As a result of the transition, there is a big jump in the number of procedure codes from about 4,000 to over 70,000.

Both ICD-10-CM and ICD-10-PCS are updated annually and are free for use without charge.

Current Procedural Terminology (CPT)

CPT is developed by the American Medical Association (AMA) to encode medical services and procedures. In the U.S., CPT is used to report physician services, many non-physician services, and surgical procedures performed in hospital outpatient departments and ambulatory surgery centers. The scope of CPT includes physician consultation and procedures, physical and occupational services, radiological and clinical laboratory investigations, transportation services, and others. There are three categories of CPT codes. Category I codes are five-digit numeric codes. For a procedure to receive a category 1 code, it must be an established and approved procedure with proven clinical efficacy performed by many healthcare professionals. Category II codes are five-character alphanumeric codes ending with an ‘F’. These are supplementary tracking codes

for quality and performance measurement. Category III codes are temporary five-character alphanumeric codes ending with 'T'. These codes are for emerging technologies that do not yet qualify for regular category I codes. There are about 9,000 category I codes. CPT is now in the fourth edition and is updated annually. Use of CPT requires a license from AMA.

Clinical data warehouses for translational research

Several clinical data warehouses have been developed for translational research purposes. On the one hand, there are traditional data warehouses created through the Clinical and Translational Science Awards (CTSA) program and other translational research efforts. Such warehouse include BTRIS [52], based on its own ontology, the Research Entity Dictionary, and STRIDE [53], based on standard ontologies, such as SNOMED CT and RxNorm. On the other hand, several proof-of-concept projects have leveraged Semantic Web technologies for translational research purposes. In the footsteps of a demonstration project illustrating the benefits of integrating data in the domain of Alzheimer's disease [54], other researchers have developed knowledge bases for cancer data (leveraging the NCI Thesaurus) [55] and in the domain of nicotine dependence (using an ontology developed specifically for the purpose of integrating publicly-available datasets) [56]. The Translational Medicine Knowledge Base, based on the Translational Ontology, is a more recent initiative developed for answering questions relating to clinical practice and pharmaceutical drug discovery [57].

Ontology repositories

Because most biomedical terminologies and ontologies are developed by different groups and institutions independently of each other and made available to users in heterogeneous formats, interoperability among them is generally limited. In order to create some level of semantic interoperability among ontologies and facilitate their use, several repositories have been created. Such repositories provide access to integrated ontologies through powerful graphical and programming interfaces. This section presents the two largest repositories: the Unified Medical Language System (UMLS) and the BioPortal.

Unified Medical Language System (UMLS)

The U.S. National Library of Medicine (NLM) started the UMLS project in 1986. One of the main goals of UMLS is to aid the development of systems that help health professionals and researchers retrieve and integrate electronic biomedical information from a multitude of disparate sources. [58-61] One major obstacle to cross-source information retrieval is that the same information is often expressed differently in different vocabularies used by the various systems and there is no universal biomedical vocabulary. Knowing that to dictate the use of a single vocabulary is not realistic, the UMLS circumvents this problem by creating links between the terms in different vocabularies. The UMLS is available free of charge. Users need to acquire a license because some of the UMLS contents are protected by additional license requirements. [62] Currently, there are over 20,000 UMLS licensees in more than 120 countries. The UMLS is released twice a year.

UMLS knowledge sources

The Metathesaurus of the UMLS is a conglomeration of a large number of terms that exist in biomedical vocabularies. All terms that refer to the same meaning (i.e. synonymous terms) are grouped together in the same UMLS concept. Each UMLS concept is assigned a permanent unique identifier (the Concept Unique Identifier, CUI), which is the unchanging pointer to that particular concept. This concept-based organization enables cross-database information retrieval based on *meaning*, independent of the lexical variability of the terms themselves. In the 2018AA release, the UMLS Metathesaurus incorporates 154 source vocabularies and includes terms in 25 languages. There are two million biomedical concepts and eight million unique terms. The Metathesaurus also contains relationships between concepts. Most of these relationships are derived from relationships asserted by the source vocabularies. To edit the Metathesaurus, the UMLS editors use a sophisticated set of lexical and rule-based matching algorithms to help them focus on areas that require manual review.

The Semantic Network is another resource in the UMLS. The Semantic Network contains 127 semantic types and 54 kinds of relationship between the semantic types. The Semantic Network is primarily used for the categorization of UMLS concepts [9]. All UMLS concepts are assigned at least one semantic type. The

semantic relationships represent the possible relationships between semantic types, which may or may not hold true at the concept level. A third resource in the UMLS is the SPECIALIST Lexicon and the lexical tools. The SPECIALIST Lexicon is a general English lexicon that includes over 500,000 lexical items. Each lexicon entry records the syntactic, morphological and orthographic information that can be used to support activities such as natural language processing of biomedical text. The lexical tools are designed to address the high degree of variability in natural language words and terms. Normalization is one of the functions of the lexical tools that helps users to abstract away from variations involving word inflection, case and word order [63].

UMLS tooling

The UMLS is distributed as a set of relational tables that can be loaded in a database management system. Alternatively, a web-based interface and an application programming interface (API) are provided. The UMLS Terminology Services (UTS) is a web-based portal that can be used for downloading UMLS data, browsing the UMLS Metathesaurus, Semantic Network and SPECIALIST Lexicon, and for accessing the UMLS documentation. Users of the UTS can enter a biomedical term or the identifier of a biomedical concept in a given ontology, and the corresponding UMLS concept will be retrieved and displayed, showing the names for this concept in various ontologies, as well as the relations of this concept to other concepts. For example, a search on “addison’s disease” retrieves all names for the corresponding concept (C0001403) in over 25 ontologies (version 2018AA, as of June 2018), including SNOMED CT, MedDRA, and the International Classification of Primary Care. Each ontology can also be navigated as a tree. In addition to the graphical interface, the UTS also offers an application programming interface (API) based on RESTful web services. This API provides access to the properties and relations of Metathesaurus concepts, as well as semantic types and lexical entries. Most functions of the UTS API require UMLS user credentials to be checked in order to gain access to UMLS data. Support for user authentication is provided through the UTS API itself.

UMLS applications

The UMLS provides convenient one-stop access to diverse biomedical vocabularies, which are updated as frequently as resources allow. One important contribution of the UMLS is that all source vocabularies are converted to a common schema of representation, with the same file structure and object model. This makes it much easier to build common tools that deal with multiple vocabularies, without the need to grapple with the native format of each. Moreover, this also enhances the understanding of the vocabularies as the common schema abstracts away from variations in naming conventions. For example, a term may be called 'preferred name', 'display name' or 'common name' in different vocabularies, but if they are determined to mean the same type of term functionally they are all referred to as 'preferred term' in the UMLS.

One common use of the UMLS is inter-terminology mapping. The UMLS concept structure enables easy identification of equivalent terms between any two source terminologies. In addition to mapping by synonymy, methods have been reported that create inter-terminology mapping by utilizing relationships and lexical resources available in the UMLS. [64] Natural language processing is another important use of the UMLS making use of its large collection of terms, the SPECIALIST Lexicon and the lexical tools. MetaMap is a publicly available tool developed by NLM which aims to identify biomedical concepts in free text. [65, 66] This is often the first step in data-mining and knowledge discovery. Other uses of the UMLS include terminology research, information indexing and retrieval, and terminology creation. [67]

BioPortal

BioBortal is developed by the National Center for Biomedical Ontology (NCBO), one of the National Centers for Biomedical Computing, created in 2004. The goal of NCBO is "to support biomedical researchers in their knowledge-intensive work, by providing online tools and a Web portal enabling them to access, review, and integrate disparate ontological resources in all aspects of biomedical investigation and clinical practice." BioPortal not only provides access to biomedical ontologies, but it also helps link ontologies to biomedical data [68].

BioPortal ontologies

The current version of BioPortal integrates over 700 ontologies for biomedicine, biology and life sciences, and includes roughly 9 million concepts. A number of ontologies integrated in the UMLS are also present in BioPortal (e.g., Gene Ontology, LOINC, NCIT, and SNOMED CT). However, BioPortal also provides access to the ontologies form the Open Biomedical Ontologies (OBO) family, an effort to create ontologies across the biomedical domain. In addition to the Gene Ontology, OBO includes ontologies for chemical entities (e.g., ChEBI), biomedical investigations (OBI), phenotypic qualities (PATO) and anatomical ontologies for several model organism, among many others. Some of these ontologies have received the "seal of approval" of the OBO Foundry (e.g., Gene Ontology, ChEBI, OBI, and Protein Ontology). Finally, the developers of biomedical ontologies can submit their resources directly to BioPortal, which makes BioPortal an open repository, as opposed to the UMLS. Examples of such resources include the Research Network and Patient Registry Inventory Ontology and the Ontology of Clinical Research. BioPortal supports several popular formats for ontologies, including OWL, OBO format and the Rich Release Format (RRF) of the UMLS.

BioPortal tooling

BioPortal is a web-based application allowing users to search, browse, navigate, visualize and comment on the biomedical ontologies integrated in its repository. For example, a search on "addison's disease" retrieves the corresponding entries in 51 ontologies (as of June 2018), including SNOMED CT, the Human Phenotype Ontology and DermLex. Visualization as tree or graph is offered for each ontology. The most original feature of BioPortal is to support the addition of marginal notes to various elements of an ontology, e.g., to propose new terms or suggest changes in relations. Such comments can be used as feedback by the developers of the ontologies and can contribute to the collaborative editing on ontologies. Users can also publish reviews of the ontologies. In addition to the graphical interface, BioPortal also offers an application programming interface (API) based on RESTful web services and is generally well integrated with Semantic Web technologies, as it provides URIs for each concept, which can be used as a reference in linked data applications.

BioPortal applications

Similar to the UMLS, BioPortal identifies equivalent concepts across ontologies in its repositories (e.g., between the term *listeriosis* in DermLex and in Medline Plus Health Topics). The BioPortal Annotator is a high-throughput named entity recognition system available both as an application and a web service. The Annotator identifies the names of biomedical concepts in text using fast string matching algorithms. While users can annotate arbitrary text, BioPortal also contains 40 million records from 50 textual resources, which have been preprocessed with the Annotator, including several gene expression data repositories, ClinicalTrials.gov and the Adverse Event Reporting System from the Food and Drug Administration (FDA). In practice, BioPortal provides an index to these resources, making it possible to use terms from its ontologies to search these resources. Finally, BioPortal also provides the Ontology Recommender, a tool that suggests the most relevant ontologies based on an excerpt from a biomedical text or a list of keywords.

Approaches to ontology alignment in ontology repositories

Apart from providing access to existing terminologies and ontologies, the UMLS and BioPortal also identify bridges between these artifacts, which will facilitate inter-ontology integration or alignment. For the UMLS, as each terminology is added or updated, every new term is comprehensively reviewed (by lexical matching followed by manual review) to see if they are synonymous with existing UMLS terms. If so, the incoming term is grouped under the same UMLS concept. In the BioPortal, equivalence between different ontologies is discovered by a different approach. For selected ontologies, possible synonymy is identified through algorithmic matching alone (without human review). It has been shown that simple lexical matching works reasonably well in mapping between some biomedical ontologies in BioPortal, compared to more advanced algorithms [69]. Users can also contribute equivalence maps between ontologies.

Ontology in action – Uses of ontologies in clinical research

Ontologies can be used to facilitate clinical research in multiple ways. In the following section, we shall highlight three areas for discussion: research workflow management, data integration and electronic phenotyping. However,

these are not meant to be water-tight categories (e.g. the ontological modeling of the research design can facilitate workflow management, as well as data sharing and integration).

Research workflow management

In most clinical trials, knowledge about protocols, assays and specimen flow is stored and shared in textual documents and spreadsheets. The descriptors used are neither encoded nor standardized. Standalone computer applications are often used to automate specific portions of the research activity (e.g. trial authoring tools, operational plan builders, study site management software). These applications are largely independent and rarely communicate with each other. Integration of these systems will result in more efficient workflow management, improve the quality of the data collected and simplify subsequent data analysis. However, the lack of common terminology and semantics to describe the characteristics of a clinical trial impedes efforts of integration. Ontology-based integration of clinical trials management applications is an attractive approach. One early example is the Immune Tolerance Network, a large distributed research consortium engaged in the discovery of new therapy for immune-related disorders. The Network created the Epoch Clinical Trial Ontologies and built an ontology-based architecture to allow sharing of information between disparate clinical trial software applications. [70] Based on the ontologies, a clinical trial authoring tool had also been developed. [71]

Another notable effort in the use of ontology in the design and implementation of clinical trials is the Advancing Clinical Genomic Trials on Cancer (ACGT) Project in Europe. ACGT is a European Union co-funded project that aims at developing open-source, semantic and grid-based technologies in support of post genomic clinical trials in cancer research. One component of this project is the development of a tool called Ontology-based Trial Management Application (ObTiMA), which has two main components: the Trial Builder and the Patient Data Management System, which are based on their master ontology called ACGT Master Ontology (ACGT-MO). [72-75] Trial Builder is used to create ontology-based case report forms (CRF) and the Patient Data Management System facilitates data collection by front-line clinicians.

The advantage of an ontology-based approach in data capture is that the alignment of research semantics and data definition is achieved early in the research process, which facilitates greatly the downstream integration of data collected from different data sources. The early use of a common master ontology obviates the need of a *post hoc* mapping between different data and information models, which is time-consuming and error-prone. Similar examples can be found in the use of OBI and BRIDG. OBI is used to define a standard submission form for the Eukaryotic Pathogen Database project, which integrates genomic and functional genomics data for over 30 protozoan parasites. [76] While the specific terms used for a specimen are mainly drawn from other ontologies (e.g., Gazetteer, PATO), OBI is used to provide categories for the terms used (e.g., sequence data) to facilitate the loading of the data onto a database and subsequent data mining. In the U.S., FDA has used the BRIDG as the conceptual model for the Janus Clinical Trials Repository (CTR) warehouse. To support drug marketing application, clinical trial sponsors need to submit subject-level data from trials in the CDISC format to the FDA for storage in the Janus CTR, which is used to support regulatory review and cross-study analysis. [77]

Data integration

In the post-genomic era of research, the power and potential value of linking data from disparate sources is increasingly recognized. A rapidly developing branch of translational research exploits the automated discovery of association between clinical and genomics data. [78] Ontologies can play important roles at different strategic steps of data integration. [79]

For many existing data sources, data sharing and integration only occurs as an after-thought. To align multiple data sources to support activities such as cross-study querying or data-mining is no trivial task. The classical approach, warehousing, is to align the sources at the **data** level (i.e. to annotate or index all available data by a common ontology). When the source data are encoded in different vocabularies or coding systems, which is sadly a common scenario, data

integration requires alignment or mapping between the vocabularies. Resources like the UMLS and BioPortal are very useful in such mapping activity.

Another approach to data integration is to align data sources at the **metadata** level, which allows effective cross database queries without actually pooling data in a common database or warehouse. The prerequisite to the effective query of a network of federated research data sources is a standard way to describe the characteristics of the individual sources. This is the role of a common research metadata ontology. OCRe (described above) is specifically created to annotate and align clinical trials according to their design and data analysis methodology. In a pilot study, OCRe is used to develop an end-to-end informatics infrastructure that enables data acquisition, logical curation and federated querying of human studies to answer questions such as "find all placebo-controlled trials in which a macrolide is used as an intervention". [28] Using similar approaches for data discovery and sharing, a brand-new platform called Vivli is created to promote the reuse of clinical research data. [80] Vivli is intended to act as a neutral broker between data contributor, data user and the wider data sharing community. It will provide an independent data repository, in-depth search engine and a cloud-based, secure analytics platform.

Another notable effort is BIRNLex which is created to annotate the Biomedical Informatics Research Network (BIRN) data sources. [56] The BIRN sources include image databases ranging from magnetic resonance imaging of human subjects, mouse models of human neurologic disease to electron microscopic imaging. BIRNLex not only covers terms in neuroanatomy, molecular species and cognitive processes, it also covers concepts such as experimental design, data types and data provenance. BIRN employs a mediator architecture to link multiple databases. The mediator integrates the various source databases by the use of a common ontology. The user query is parsed by the mediator, which issues database-specific queries to the relevant data sources each with their specific local schema. [81]

The use of OBI in the Investigation, Study, Assay (ISA) Project is another example of ontology-based facilitation of data integration and sharing. The ISA

Project supports managing and tracking biological experiment metadata to ensure its preservation, discoverability and re-use. [82] Concepts from OBI are used to annotate the experimental design and other characteristics, so that queries such as "retrieve all studies with balanced design" or "retrieve all studies where study groups have at least 3 samples" are possible. In a similar vein, the BRIDG model ontology is used in various projects to facilitate data exchange. One example is the SALUS (Security and interoperability in next generation Public Protection and Disaster Relief (PPDR) communication infrastructures) Project of the European Union. [83] BRIDG is used to provide semantics for the project's metadata repository to allow meaningful exchange of data between European electronic health records.

Other innovative approaches of using ontologies to achieve data integration have also been described. One study explored the possibility of tagging research data to support real-time meta-analysis. [84] Another described a prototype system for ontology-driven indexing of public data sets for translational research. [85]

One particular form of data integration supported by ontologies is represented by what has become known as "Linked Data" in the Semantic Web community [86]. The foundational idea behind linked data and the Semantic Web is that resources semantically annotated to ontologies can be interrelated when they refer to the same entities. In practice, datasets are represented as graphs in RDF, the Resource Description Framework, in which nodes (representing entities) can be shared across graphs, enabling connections among graphs. Interestingly, a significant portion of the datasets currently interrelated as Linked Data consists of biomedical resources, including PubMed, KEGG and DrugBank. For privacy reasons, very few clinical datasets have been made publicly available, and no such datasets are available as Linked Data yet. However, researchers have illustrated the benefits of Semantic Web technologies for translational research [54-57]. Moreover, the development of personal health records will enable individuals to share their clinical data and effective de-identification techniques might also contribute to the availability of clinical data, which could enable knowledge discovery through the mining of large volume of data. Ontologies support Linked Data in three important ways. Ontologies provide a controlled vocabulary for entities in the

Semantic Web; integrated ontology repositories, such as the UMLS and BioPortal, support the reconciliation of entities annotated to different ontologies; finally, relations in ontologies can be used for subsumption and other kinds of reasoning. An active community of researchers is exploring various aspects of biomedical linked data as part of the Semantic Web Health Care and Life Sciences interest group [87], with particular interest in the domain of drug discovery through the Linking Open Drug Data initiative [88].

Electronic phenotyping

Data in electronic health records (EHRs) are becoming increasingly available for clinical and translational research. Through projects such as the Electronic Medical Records and Genomics (eMERGE) Network, [89] National Patient-Centered Clinical Research Network (PCORnet), [90] Strategic Health IT Advanced Research Projects (SHARP), [91] Observational Health Data Sciences and Informatics (OHDSI), [92] and NIH Health Care Systems Collaboratory, [93] it has been demonstrated that EHR data can be used to develop research-grade disease phenotypes with sufficient accuracy to identify traits and diseases for biomedical research and clinical care.

Electronic phenotyping refers to activities and applications that use data captured in the delivery of healthcare (typically from EHRs and insurance claims) to identify individuals or populations (cohorts) with clinical characteristics, events or service patterns that are relevant to interventional, observational, prospective, and/or retrospective studies. [93] So far, the most tried-and-true approach to electronic phenotyping utilizes explicit, standardized queries – consisting of logical operators, data fields and list of codes often from standardized terminologies – that can be run against different data sources to identify comparable populations. Due to the heterogeneity across care settings, data models and patient populations, designing phenotype definitions is complex and often requires customization for different data sources. However, the validity of selected phenotype definitions and the comparability of patient populations across different health care settings has been reported. [94-98] Newer approaches in electronic phenotyping involving techniques such as machine learning have been

studied with promising results. [99-102] However, manually curated phenotype definitions are still the most commonly employed phenotyping method.

Most phenotype definitions to date use both structured and unstructured elements in the EHR. Structured elements usually include demographic information, billing codes, laboratory tests, vital signs and medications. Unstructured elements include clinical notes, family history, radiology reports, pathology reports and others. Utilization of unstructured data elements usually require additional processing by natural language processing. So far, the most commonly used structured data are the billing codes – especially the ICD and CPT codes because of their ubiquity in the EHR. [103] With the increasing use of clinical terminologies such as SNOMED CT, LOINC and RxNorm as a result of the Meaningful Use and subsequent incentive programs, it is expected that the inclusion of these terminologies in phenotype definitions will increase. This should have a positive impact in the accuracy of phenotyping as clinical terminologies such as SNOMED CT have been shown to provide better coverage and more fine-grained representation of clinical information. [104-106] The use of standardized terminologies in the EHR will be a great boon towards making phenotype definitions fully computable and portable across data sources. [107] The use of robust terminologies can also make phenotype authoring more efficient. For example, the tools developed by the Informatics for Integrating Biology and the Bedside (i2b2) project leverage the intrinsic hierarchical structure of medical ontologies to allow the selection of all descendants under the same concept. [108] Before standardized terminologies become the norm, the diversity in content terminologies remains a challenge to electronic phenotyping. One approach to mitigate this problem is demonstrated by the OHDSI collaborative. The OHDSI vocabulary incorporates and maps terms from different terminologies to a core list of concepts.

Development of phenotype definitions is a time and resource intensive activity. Often knowledge engineers, domain experts and researchers have to spend many hours to create and iteratively refine phenotype algorithms to achieve high sensitivity, specificity, positive and negative predictive values. It is highly likely that different research groups have the need to identify some common conditions

such as type 2 diabetes mellitus. To ensure comparability of results and to avoid duplication of effort, it is important that phenotype definitions are validated and shared across institutional and organizational boundaries. One platform for the creation, validation and dissemination of phenotype definitions is the Phenotype Knowledgebase (PheKB) developed by the eMERGE Network. [103] PheKB has built-in tools specifically designed to enhance knowledge sharing and collaboration, so as to facilitate the transportability of phenotype definitions across different health care systems, clinical data repositories and research applications.

Phenotype definitions often include enumerated lists of concepts that identify the pertinent characteristics of a patient population. These lists are conventionally called value sets, which are lists of codes from standard terminologies for diagnosis, procedures, laboratory tests, medications etc. Value sets developed for phenotype definitions are very similar to value sets developed for quality measures. As part of the Electronic Clinical Quality Improvement (eCQI) initiative, health care systems have to submit data for selected clinical quality measures. [109] Quality measure value sets are used to identify sub-populations of patients sharing certain demographic and clinical characteristics, as defined by a clinical quality measure. The Value Set Authority Center (VSAC) of NLM is a purpose-built platform to support the authoring, maintenance and dissemination of value sets which can be used for quality measurement, phenotype definition and other purposes. [110]

The Way Forward

Looking forward, it is encouraging that the value of ontologies in clinical research becomes more recognized. This is evidenced by the increase in the number of investigations making use of ontologies. At the same time, this is also accompanied by an increase in the number of ontologies, which in itself is a mixed blessing. Many researchers still tend to create their own ontologies to suit their specific use case. Re-use of existing ontologies is only a rarity. If left unchecked, this tendency has the potential of growing into the very problem that ontologies are created to solve – the multitude of ontologies will itself become the barrier to data interoperability and integration. *Post hoc* mapping and alignment of

ontologies is often difficult (if not impossible) and an approximation at best (with inherent information loss). The solution is to coordinate the development and maximize the re-use of existing ontologies, which will significantly simplify things downstream.

To facilitate re-use of ontologies, resources like the UMLS and BioPortal are indispensable. They enable users to navigate the expanding sea of biomedical ontologies. In addition to listing and making these ontologies available, what is still lacking is a better characterization of these ontologies to help users decide whether they are suitable for the tasks at hand. In case there are multiple candidate ontologies, some indicators of quality (e.g. user base, ways in which they are used, user feedback and comments) will be very useful to help users decide on the best choice.

Acknowledgments

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM).

References

1. Bodenreider, O., *Biomedical ontologies in action: role in knowledge management, data integration and decision support.* Yearb Med Inform, 2008: p. 67-79.
2. Smith, B., *Ontology (Science).* Nature Precedings, 2008. **Available from Nature Precedings (http://hdl.handle.net/10101/npre.2008.2027.2)**.
3. Bodenreider, O. and R. Stevens, *Bio-ontologies: current trends and future directions.* Brief Bioinform, 2006. **7**(3): p. 256-74.
4. Cimino, J.J. and X. Zhu, *The practical impact of ontologies on biomedical informatics.* Yearb Med Inform, 2006: p. 124-35.
5. Smith, B., et al., *Relations in biomedical ontologies.* Genome Biol, 2005. **6**(5): p. R46.
6. Simmons, P. and J. Melia, *Continuants and occurrents.* Proceedings of the Aristotelian Society, Supplementary Volumes, 2000. **74**: p. 59-75+77-92.
7. IFOMIS. *BFO*. Available from: http://www.ifomis.org/bfo/.
8. Laboratory for Applied Ontology. *DOLCE*. Available from: http://www.loa-cnr.it/DOLCE.html.
9. McCray, A.T., *An upper-level ontology for the biomedical domain.* Comp Funct Genomics, 2003. **4**(1): p. 80-4.
10. Baader, F., et al., eds. *The description logic handbook : theory, implementation, and applications*. 2nd ed.

xix, 601 p ed. 2007, Cambridge University Press: Cambridge New York. ill. 26 cm.
11. Berners-Lee, T., J. Hendler, and O. Lassila, *The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities.* Scientific American, 2001. **284**(5): p. 34-43.
12. World Wide Web Consortium *OWL 2 Web Ontology Language Document Overview*. 2009; Available from: http://www.w3.org/TR/owl2-overview/.
13. World Wide Web Consortium *RDF Vocabulary Description Language 1.0: RDF Schema*. 2004; Available from: http://www.w3.org/TR/rdf-schema/.
14. World Wide Web Consortium *SKOS Simple Knowledge Organization System Reference*. 2009; Available from: http://www.w3.org/TR/2009/REC-skos-reference-20090818/.
15. Mungall, C., et al. *OBO Flat File Format 1.4 Syntax and Semantics*. Available from: http://owlcollab.github.io/oboformat/doc/obo-syntax.html.
16. Day-Richter, J. *The OBO Flat File Format Specification*. 2006; Available from: http://www.geneontology.org/GO.format.obo-1_2.shtml.
17. Golbreich, C., et al., *OBO and OWL: leveraging semantic web technologies for the life sciences*, in *Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference*. 2007, Springer-Verlag: Busan, Korea. p. 169-182.
18. Noy, N., et al., *The ontology life cycle: Integrated tools for editing, publishing, peer review, and evolution of ontologies.* AMIA Annu Symp Proc, 2010. **2010**: p. 552-6.
19. Stanford Center for Biomedical Informatics Research. *Protégé*. Available from: http://protege.stanford.edu/.
20. Day-Richter, J., et al., *OBO-Edit--an ontology editor for biologists.* Bioinformatics, 2007. **23**(16): p. 2198-200.
21. Lawrence Berkeley National Lab. *OBO-Edit*. Available from: http://oboedit.org/.
22. Smith, B., et al., *The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration.* Nat Biotechnol, 2007. **25**(11): p. 1251-5.
23. International, S. *Partnerships -- Working with other Standards Organizations*. Available from: https://www.snomed.org/about/partnerships.
24. Richesson, R.L. and J. Krischer, *Data standards in clinical research: gaps, overlaps, challenges and future directions.* J Am Med Inform Assoc, 2007. **14**(6): p. 687-96.
25. McQuilton P., G.-B.A., Rocca-Serra P., Thurston M., Lister A., Maguire E., Sansone SA., *BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences.* Database (Oxford), 2016 May 17.
26. *FAIRsharing website: https://www.FAIRsharing.org*.
27. Tu, S.W., et al., *OCRe: Ontology of Clinical Research*, in *11th International Protege Conference*. 2009.
28. Sim, I., et al., *Ontology-based federated data access to human studies information.* AMIA Annu Symp Proc, 2012. **2012**: p. 856-65.
29. *Ontology for Biomedical Investigations: Community Standard for Scientific Data Integration*. Available from: http://obi-ontology.org/.

30. Bandrowski, A., et al., *The Ontology for Biomedical Investigations.* PLoS One, 2016. **11**(4): p. e0154556.
31. Whetzel, P.L., et al., *Development of FuGO: an ontology for functional genomics investigations.* Omics, 2006. **10**(2): p. 199-204.
32. Brinkman, R.R., et al., *Modeling biomedical experimental processes with OBI.* J Biomed Semantics, 2010. **1 Suppl 1**: p. S7.
33. *Biomedical Research Integrated Domain Group Website*. Available from: https://bridgmodel.nci.nih.gov/faq/components-of-bridg-model.
34. Fridsma, D.B., et al., *The BRIDG Project: A Technical Report.* J Am Med Inform Assoc, 2008. **15**(2): p. 130-137.
35. Becnel, L.B., et al., *BRIDG: a domain information model for translational and clinical protocol-driven research.* J Am Med Inform Assoc, 2017. **24**(5): p. 882-890.
36. Tu, S.W., et al., *Bridging Epoch: Mapping Two Clinical Trial Ontologies*, in *10th International Protege Conference*. 2007.
37. de Coronado, S., et al., *NCI Thesaurus: using science-based terminology to integrate cancer research results.* Medinfo, 2004. **11**(Pt 1): p. 33-7.
38. Sioutos, N., et al., *NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information.* J Biomed Inform, 2007. **40**(1): p. 30-43.
39. Fragoso, G., et al., *Overview and Utilization of the NCI Thesaurus.* Comp Funct Genomics, 2004. **5**(8): p. 648-54.
40. International, S. *SNOMED CT (Systematized Nomenclature of Medicine-Clinical Terms), SNOMED International*. Available from: https://www.snomed.org/.
41. Lee, D., et al., *A survey of SNOMED CT implementations.* J Biomed Inform, 2013. **46**(1): p. 87-96.
42. Blumenthal, D. and M. Tavenner, *The "meaningful use" regulation for electronic health records.* N Engl J Med, 2010. **363**(6): p. 501-4.
43. Office of the National Coordinator for Health Information Technology (ONC) - Department of Health and Human Services, *Standards & Certification Criteria Interim Final Rule: Revisions to Initial Set of Standards, Implementation Specifications, and Certification Criteria for Electronic Health Record Technology.* Federal Register, 2010. **75**(197): p. 62686-62690.
44. Huff, S.M., et al., *Development of the Logical Observation Identifiers Names and Codes (LOINC) vocabulary.* J Am Med Inform Assoc, 1998. **5**(3): p. 276-92.
45. *Logical Observation Identifier Names and Codes (LOINC)*. Available from: https://loinc.org/.
46. Nelson, S.J., et al., *Normalized names for clinical drugs: RxNorm at 6 years.* J Am Med Inform Assoc, 2011. **18**(4): p. 441-8.
47. Bouhaddou, O., et al., *Exchange of Computable Patient Data Between the Department of Veterans Affairs (VA) and the Department of Defense (DoD): Terminology Standards Strategy.* J Am Med Inform Assoc, 2007.
48. *History of the development of the ICD, World Health Organization*. Available from: http://www.who.int/classifications/icd/en/HistoryOfICD.pdf.
49. Steindel, S.J., *International classification of diseases, 10th edition, clinical modification and procedure coding system: descriptive overview of the*

next generation HIPAA code sets. J Am Med Inform Assoc, 2010. **17**(3): p. 274-82.
50. Fung, K.W., et al., *Preparing for the ICD-10-CM Transition: Automated Methods for Translating ICD Codes in Clinical Phenotype Definitions.* EGEMS (Wash DC), 2016. **4**(1): p. 1211.
51. Averill, R.F., et al., *Development of the ICD-10 procedure coding system (ICD-10-PCS).* Top Health Inf Manage, 2001. **21**(3): p. 54-88.
52. Cimino, J.J. and E.J. Ayres, *The clinical research data repository of the US National Institutes of Health.* Stud Health Technol Inform, 2010. **160**(Pt 2): p. 1299-303.
53. Lowe, H.J., et al., *STRIDE--An integrated standards-based translational research informatics platform.* AMIA Annu Symp Proc, 2009. **2009**: p. 391-5.
54. Ruttenberg, A., et al., *Methodology - Advancing translational research with the Semantic Web.* Bmc Bioinformatics, 2007. **8**: p. -.
55. McCusker, J.P., et al., *Semantic web data warehousing for caGrid.* Bmc Bioinformatics, 2009. **10 Suppl 10**: p. S2.
56. Sahoo, S.S., et al., *An ontology-driven semantic mashup of gene and biological pathway information: application to the domain of nicotine dependence.* J Biomed Inform, 2008. **41**(5): p. 752-65.
57. Semantic Web for Health Care and Life Sciences Interest Group. *Translational Medicine Ontology and Knowledge Base*. Available from: http://www.w3.org/wiki/HCLSIG/PharmaOntology.
58. Humphreys, B.L., D.A. Lindberg, and W.T. Hole, *Assessing and enhancing the value of the UMLS Knowledge Sources.* Proc Annu Symp Comput Appl Med Care, 1991: p. 78-82.
59. Humphreys, B.L., et al., *The Unified Medical Language System: an informatics research collaboration.* J Am Med Inform Assoc, 1998. **5**(1): p. 1-11.
60. Lindberg, D.A., B.L. Humphreys, and A.T. McCray, *The Unified Medical Language System.* Methods Inf Med, 1993. **32**(4): p. 281-91.
61. Bodenreider, O., *The Unified Medical Language System (UMLS): Integrating biomedical terminology.* Nucleic Acids Res, 2004. **32 Database issue**: p. D267-70.
62. UMLS. *Unified Medical Language System (UMLS)*. Available from: http://www.nlm.nih.gov/research/umls/.
63. McCray, A.T., S. Srinivasan, and A.C. Browne, *Lexical methods for managing variation in biomedical terminologies.* Proc Annu Symp Comput Appl Med Care, 1994: p. 235-9.
64. Fung, K.W. and O. Bodenreider, *Utilizing the UMLS for semantic mapping between terminologies.* AMIA Annu Symp Proc, 2005: p. 266-70.
65. Aronson, A.R., *Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.* Proc AMIA Symp, 2001: p. 17-21.
66. Aronson, A.R. and F.M. Lang, *An overview of MetaMap: historical perspective and recent advances.* J Am Med Inform Assoc, 2010. **17**(3): p. 229-36.
67. Fung, K.W., W.T. Hole, and S. Srinivasan, *Who is using the UMLS and how - insights from the UMLS user annual reports.* AMIA Annu Symp Proc, 2006: p. 274-8.

68. Noy, N.F., et al., *BioPortal: ontologies and integrated data resources at the click of a mouse.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W170-3.
69. Ghazvinian, A., N.F. Noy, and M.A. Musen, *Creating mappings for ontologies in biomedicine: simple methods work.* AMIA Annu Symp Proc, 2009. **2009**: p. 198-202.
70. Shankar, R.D., et al., *An ontology-based architecture for integration of clinical trials management applications.* AMIA Annu Symp Proc, 2007: p. 661-5.
71. Shankar, R., et al., *TrialWiz: an Ontology-Driven Tool for Authoring Clinical Trial Protocols.* AMIA Annu Symp Proc, 2008: p. 1226.
72. Margolin, A.A., et al., *ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context* 2006, Columbia University.
73. Brochhausen, M., et al., *The ACGT Master Ontology and its applications--towards an ontology-driven cancer research and management system.* J Biomed Inform, 2011. **44**(1): p. 8-25.
74. Weiler, G., et al., *Ontology based data management systems for post-genomic clinical trials within a European Grid Infrastructure for Cancer Research.* Conf Proc IEEE Eng Med Biol Soc, 2007. **2007**: p. 6435-8.
75. Stenzhorn, H., et al., *The ObTiMA system - ontology-based managing of clinical trials.* Stud Health Technol Inform, 2010. **160**(Pt 2): p. 1090-4.
76. *Eukaryotic Pathogen Database*. Available from: https://eupathdb.org/eupathdb/.
77. *FDA Janus Data Repository*. Available from: https://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/ucm155327.htm.
78. *Genome-Wide Association Studies*. Available from: http://grants.nih.gov/grants/gwas/.
79. Bodenreider, O., *Ontologies and data integration in biomedicine: Success stories and challenging issues*, in *Proceedings of the Fifth International Workshop on Data Integration in the Life Sciences (DILS 2008)*, A. Bairoch, S. Cohen-Boulakia, and C. Froidevaux, Editors. 2008, Springer: Berlin Heidelberg New York. p. 1-4.
80. *Vivli: Center for Global Clinical Research Data*. Available from: http://vivli.org/.
81. Rubin, D.L., N.H. Shah, and N.F. Noy, *Biomedical ontologies: a functional perspective.* Brief Bioinform, 2008. **9**(1): p. 75-90.
82. Sansone, S.A., et al., *Toward interoperable bioscience data.* Nat Genet, 2012. **44**(2): p. 121-6.
83. *SALUS Project: Security and interoperability in next generation PPDR communication infrastructures*. Available from: https://www.sec-salus.eu/.
84. Cook, C., et al., *Real-time updates of meta-analyses of HIV treatments supported by a biomedical ontology.* Account Res, 2007. **14**(1): p. 1-18.
85. Shah, N.H., et al., *Ontology-driven indexing of public datasets for translational bioinformatics.* BMC Bioinformatics, 2009. **10 Suppl 2**: p. S1.
86. Bizer, C., T. Heath, and T. Berners-Lee, *Linked Data - The Story So Far.* International Journal on Semantic Web and Information Systems, 2009. **5**(3): p. 1-22.

87. HCLS, *Semantic Web Health Care and Life Sciences (HCLS) Interest Group.*
88. Semantic Web for Health Care and Life Sciences Interest Group. *Linking Open Drug Data*. Available from: http://www.w3.org/wiki/HCLSIG/LODD.
89. Gottesman, O., et al., *The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future.* Genet Med, 2013. **15**(10): p. 761-71.
90. Fleurence, R.L., et al., *Launching PCORnet, a national patient-centered clinical research network.* J Am Med Inform Assoc, 2014. **21**(4): p. 578-82.
91. Chute, C.G., et al., *The SHARPn project on secondary use of Electronic Medical Record data: progress, plans, and possibilities.* AMIA Annu Symp Proc, 2011. **2011**: p. 248-56.
92. Hripcsak, G., et al., *Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers.* Stud Health Technol Inform, 2015. **216**: p. 574-8.
93. Richesson, R.L., et al., *Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory.* J Am Med Inform Assoc, 2013. **20**(e2): p. e226-31.
94. Newton, K.M., et al., *Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network.* J Am Med Inform Assoc, 2013. **20**(e1): p. e147-54.
95. Ritchie, M.D., et al., *Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record.* Am J Hum Genet, 2010. **86**(4): p. 560-72.
96. Kho, A.N., et al., *Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study.* J Am Med Inform Assoc, 2012. **19**(2): p. 212-8.
97. Carroll, R.J., et al., *Portability of an algorithm to identify rheumatoid arthritis in electronic health records.* J Am Med Inform Assoc, 2012. **19**(e1): p. e162-9.
98. Cutrona, S.L., et al., *Validation of acute myocardial infarction in the Food and Drug Administration's Mini-Sentinel program.* Pharmacoepidemiol Drug Saf, 2013. **22**(1): p. 40-54.
99. Hripcsak, G. and D.J. Albers, *Next-generation phenotyping of electronic health records.* J Am Med Inform Assoc, 2013. **20**(1): p. 117-21.
100. Martin-Sanchez, F.J., et al., *Secondary Use and Analysis of Big Data Collected for Patient Care.* Yearb Med Inform, 2017. **26**(1): p. 28-37.
101. Yu, S., et al., *Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources.* J Am Med Inform Assoc, 2015. **22**(5): p. 993-1000.
102. Banda, J.M., et al., *Electronic phenotyping with APHRODITE and the Observational Health Sciences and Informatics (OHDSI) data network.* AMIA Jt Summits Transl Sci Proc, 2017. **2017**: p. 48-57.
103. Kirby, J.C., et al., *PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability.* J Am Med Inform Assoc, 2016. **23**(6): p. 1046-1052.

104. Campbell, J.R. and T.H. Payne, *A comparison of four schemes for codification of problem lists.* Proc Annu Symp Comput Appl Med Care, 1994: p. 201-5.
105. Chute, C.G., et al., *The content coverage of clinical classifications. For The Computer-Based Patient Record Institute's Work Group on Codes & Structures.* J Am Med Inform Assoc, 1996. **3**(3): p. 224-33.
106. Campbell, J.R., et al., *Phase II evaluation of clinical coding schemes: completeness, taxonomy, mapping, definitions, and clarity. CPRI Work Group on Codes and Structures.* J Am Med Inform Assoc, 1997. **4**(3): p. 238-51.
107. Mo, H., et al., *Desiderata for computable representations of electronic health records-driven phenotype algorithms.* J Am Med Inform Assoc, 2015. **22**(6): p. 1220-30.
108. Murphy, S.N., et al., *Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2).* J Am Med Inform Assoc, 2010. **17**(2): p. 124-30.
109. *Electronic Clinical Quality Improvement Resource Center, The Office of the National Coordinator for Health Information Technology*. Available from: https://ecqi.healthit.gov/content/about-ecqi.
110. *Value Set Authority Center, National Library of Medicine* Available from: https://vsac.nlm.nih.gov/.